Title Page

Title

Assessing the credibility of findings from non-randomised studies of interventions

Authors and affiliations

Luke A McGuinness, MSc [1]

Julian PT Higgins, PhD [1,2]

Jonathan AC Sterne, PhD [1,2]

[1] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

[2] NIHR Bristol Biomedical Research Centre, University of Bristol, Bristol, UK

Corresponding Author

Professor Julian PT Higgins

Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, United Kingdom

julian.higgins@bristol.ac.uk

+44 (0) 117 3314529

Second revision date

8th June 2018

Word Count

1219

Article type

Viewpoint

Manuscript Subject Area

Statistics and research methods

Assessing the credibility of findings from non-randomised studies of interventions

Randomised controlled trials (RCTs) are widely accepted as the gold standard in assessment of health interventions. However, evidence from RCTs is often unavailable. They may be impractical, for example when the outcome of interest is so rare, or long-term, that appropriately-sized studies would be too expensive. RCTs require clinical equipoise, and withholding established treatments may be considered unethical even if evidence for their effects is limited. Or, RCTs may be underway but not due to report for some years.

In such situations, evidence from non-randomised studies of interventions, such as cohort studies and case-control studies, can be essential in assessing the effects of an intervention. Moreover, the increasing availability of large electronic health record datasets is leading to unprecedented opportunities to compare the effects of healthcare interventions being used in routine practice. For example, a recent study published in this journal examined the effects of multiple arterial grafting versus standard treatment on rare safety events and long-term benefit. RCTs are often unable to examine rare and long-term outcomes, and so data on over 20000 patients from the Cardiac Services British Columbia registry were analysed.[1]

Observational studies have inherent limitations, including biases due to confounding (when prognostic factors affect the intervention received), participant selection, missing data (on confounders or outcomes), and outcome measurement error. Because such biases may be unfamiliar or hard to identify, it is essential to adopt a structured approach when considering the extent to which they are present when evaluating evidence from non-randomised studies.

Formal evaluations of risk of bias are frequently undertaken in the context of a systematic review, although they are sometimes undertaken for other purposes. Numerous tools have been developed for assessing the methodological quality of observational studies. One of the best and most commonly used is the Newcastle-Ottawa Scale,[2] although this overlooks some important sources of bias. Our team recently proposed a new tool, the Risk Of Bias In

Non-randomised Studies of Interventions (ROBINS-I) tool, based on modern ideas of causal inference in epidemiology.[3] Developed through expert consensus and extensive piloting, ROBINS-I follows a recent shift in focus from *research quality* to *risk of bias* and directly addresses a study's internal validity.

ROBINS-I considers each non-randomised study as an attempt to mimic a 'target trial', defined as a large, pragmatic, randomised trial that assesses the effect of the same intervention in the same population, but which does not have factors putting it at risk of bias. Bias is then defined as the non-randomised study's tendency to produce results that differ systematically from those produced by the target trial. The hypothetical target trial does not need to be feasible or ethical: for example, it might involve randomisation of patients to intensive care versus a bed on a standard ward.

Based on the target trial and the effect of interest, ROBINS-I assesses the potential for bias to be introduced into a study's result across seven bias domains. The first, and most widely appreciated, is bias due to confounding, which arises from differences in the distribution of prognostic factors between intervention groups at the start of intervention. This is the main bias that randomisation seeks to overcome. For example a review of randomised and non-randomised studies of intra-aortic balloon pump therapy for acute myocardial infarction found little evidence of benefit in the randomised trials but substantial variation between results of non-randomised studies.[4] Much of this variation seemed to be explained by differences between characteristics of intervention groups at baseline, suggesting that confounding was introducing bias.

The second domain is bias in selection of participants into the study, which considers the extent to which the association between intervention and outcome may have been distorted due to selecting participants based on events occurring after the start of intervention. The third is bias in classification of interventions, which includes consideration of whether information about intervention status might have been affected by subsequent outcomes (including the well-known problem of 'recall bias' in case-control studies).

The remaining domains relate to biases that can also affect results of randomised trials. The fourth is bias due to deviations from intended interventions, which is particularly important when interest focusses on the effect of adherence to intervention (for which deviations clearly introduce problems). The fifth is bias due to missing data, and the sixth is bias in measurement of outcomes. The last domain, bias in the selection of the reported results, tackles the rather different problem of authors presenting only the most exciting (or statistically significant) findings.

Within each domain, ROBINS-I poses a series of broadly factual questions ('signalling questions') about what happened in the study, the answers to which directly inform a risk-of-bias judgement for that domain. In turn, the domain-level judgements inform an overall judgement (of 'low', 'moderate', 'serious' or 'critical' risk of bias) for the result in question.

Specifying a target trial can help reveal important biases in non-randomised studies that may not be immediately apparent. For example, consider the well-known erroneous finding from the Nurses' Health observational cohort study that hormone replacement therapy (HRT) is associated with an approximate halving of cardiovascular risk.[5] A hypothetical trial of HRT would randomise eligible women not currently taking HRT to start either HRT or placebo. In contrast, the Nurses' Health Study analysis included women who had started taking HRT before follow-up began, leading to selection bias. Re-analysis of the Nurses' Health Study using an approach that mimicked a randomised trial produced results similar to that from a large trial.[6] On the other hand, in an analysis of primary care electronic medical records that mimicked a target trial of statins and antihypertensives compared with no treatment, unmeasured confounding was a likely explanation for underestimation of the benefit of treatment.[7]

Specifying a target trial requires articulation of the population of interest, the experimental and comparator interventions, and the outcome(s) of interest. This creates a need for interdisciplinary collaboration: input from both methodological and subject-specific experts is necessary to define the target trial and identify the participant characteristics that might

introduce confounding. The ROBINS-I tool also requires a decision as to whether interest focusses on the effect of being assigned to an intervention (often termed the 'intention-to-treat' effect) or the effect of adhering to the assigned intervention (often termed the 'per-protocol' effect) and assesses risk of bias in relation to the chosen effect of interest.

Due to the assessment's foundation in a target randomised trial, a rating of low risk of bias would mean that the results may be regarded as equivalent to those from a high quality randomised trial. Because it is usually impossible to exclude residual or unmeasured confounding, we expect most results from non-randomised studies to be assessed at best as at moderate risk of bias. A rating of critical risk of bias implies that the result should not be included in any quantitative evidence synthesis.

The ROBINS-I tool is being adopted widely, including by Cochrane and the GRADE Working Group.[8] We expect its uptake within the context of systematic reviews and meta-analyses to increase substantially, helping to avoid review authors over-interpreting findings from non-randomised evidence, and facilitating explanations of discrepancies between findings from randomised and non-randomised studies. We also hope that the tool offers a useful framework for funders, peer reviewers and readers of journals such as *JAMA Cardiology* to think through the extent to which findings from a non-randomised study are credible.

Resources

The ROBINS-I tool, along with details of the development team and extensive guidance, is available at www.riskofbias.info.

Additional Sections

Funding/Support

LAM is funded by a National Institute for Health Research (NIHR) Systematic Review Fellowship (NIHR-RM-SR-2016-07-26). JTPH and JACS are funded by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol and by NIHR Senior Investigator awards NF-SI-0611-10168 and NF-SI-0617-10145 respectively. They are members of the NIHR Collaboration for Leadership in Applied Health Research and Care West (CLAHRC West) at University Hospitals Bristol NHS Foundation Trust, and the MRC Integrative Epidemiology Unit at the University of Bristol. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the Department of Health or the Medical Research Council. Development of ROBINS-I was funded by a Methods Innovation Fund grant from Cochrane and by Medical Research Council (MRC) grant MR/M025209/1.

Bibliography

1. Pu A, Ding L, Shin J, et al. Long-term outcomes of multiple arterial coronary artery bypass grafting: A population-based study of patients in British Columbia, Canada. *JAMA Cardiology.* 2017;2(11):1187-1196.
2. Wells GA, Shea B, O'Connell D, et al. The Newcastle-Ottawa Scale (NOS) for assessing the quality of nonrandomised studies in meta-analyses. http://www.ohri.ca/programs/clinical_epidemiology/oxford.asp. Accessed April 30, 2018.
3. Sterne JAC, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ.* 2016;355:i4919.
4. Ahmad Y, Sen S, Shun-Shin MJ, et al. Intra-aortic balloon pump therapy for acute myocardial infarction: a meta-analysis. *JAMA Internal Medicine.* 2015;175(6):931-939.
5. Stampfer MJ, Colditz GA. Estrogen replacement therapy and coronary heart disease: a quantitative assessment of the epidemiologic evidence. *Preventive Medicine.* 1991;20(1):47-63.
6. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology.* 2008;19(6):766.
7. Danaei G, Rodríguez LAG, Cantero OF, Logan RW, Hernán MA. Electronic medical records can be used to emulate target trials of sustained treatment strategies. *Journal of Clinical Epidemiology.* 2018;96:12-22.
8. Schünemann HJ, Cuello C, Akl EA, et al. GRADE Guidelines: 18. How ROBINS-I and other tools to assess risk of bias in non-randomized studies should be used to rate the certainty of a body of evidence. *Journal of Clinical Epidemiology.* Published online 9 Feb 2018.